PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/31, 15/63, 15/00, C12P 21/02

A1

(11) International Publication Number: WO 96/33276

(43) International Publication Date: 24 October 1996 (24.10.96)

(21) International Application Number: PCT/US96/05320

(22) International Filing Date: 22 April 1996 (22.04.96)

(30) Priority Data:

08/426,787	21 April 1995 (21.04.95)	US
08/476,102	7 June 1995 (07.06.95)	US
08/487,429	7 June 1995 (07.06.95)	US

(71) Applicants: HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850 (US). JOHNS HOPKINS UNIVERSITY [US/US]; 720 Rutland Avenue, Baltimore, MD 21205 (US).

(72) Inventors: FLEISCHMANN, Robert, D.; 470 Tschiffely Square Road, Gaithersburg, MD 20878 (US). ADAMS, Mark, D.; 15205 Dufief Drive, N. Potomac, MD 20878 (US). WHITE, Owen; Apartment #202, 886 Quince Orchard Boulevard, Gaithersburg, MD 20878 (US). SMITH, Hamilton, O.; 8222 Carrbridge Circle, Towson, MD 21204 (US). VENTER, J., Craig; 11915 Glen Mill Road, Potomac, MD 20854 (US).

(74) Agents: GOLDSTEIN, Jorge, A. et al.; Steme, Kessler, Goldstein & Fox P.L.L.C., Suite 600, 1100 New York Avenue, Washington, DC 20005-3934 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: NUCLEOTIDE SEQUENCE OF THE HAEMOPHILUS INFLUENZAE Rd GENOME, FRAGMENTS THEREOF, AND USES THEREOF

(57) Abstract

The present invention provides the sequencing of the entire genome of *Haemophilus influenzae* Rd, SEQ ID NO:1. The present invention further provides the sequence information stored on computer readable media, and computer-based systems and methods which facilitate its use. In addition to the entire genomic sequence, the present invention identifies over 1700 protein encoding fragments of the genome and identifies, by position relative to a unique *Not I* restriction endonuclease site, any regulatory elements which modulate the expression of the protein encoding fragments of the *Haemophilus* genome.

Even if all of the very rare sequencing errors in SEQ ID NO:1 were corrected, the resulting nucleotide sequence would still be at least 99.9% identical to the nucleotide sequence in SEQ ID NO:1.

The nucleotide sequences of the genomes from different strains of *Haemophilus influenzae* differ slightly. However, the nucleotide sequence of the genomes of all *Haemophilus influenzae* strains will be at least 99.9% identical to the nucleotide sequence provided in SEQ ID NO:1.

Thus, the present invention further provides nucleotide sequences which are at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 in a form which can be readily used, analyzed and interpreted by the skilled artisan. Methods for determining whether a nucleotide sequence is at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 are routine and readily available to the skilled artisan. For example, the well known **fasta** algothrithm (Pearson and Lipman, *Proc. Natl. Acad. Sci. USA* 85:2444 (1988)) can be used to generate the percent identity of nucleotide sequences.

Computer Related Embodiments

The nucleotide sequence provided in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 may be "provided" in a variety of mediums to facilitate use thereof. As used herein, provided refers to a manufacture, other than an isolated nucleic acid molecule, which contains a nucleotide sequence of the present invention, i.e., the nucleotide sequence provided in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1. Such a manufacture provides the *Haemophilus influenzae* Rd genome or a subset thereof (e.g., a *Haemophilus Influenzae* Rd open reading frame (ORF)) in a form which allows a skilled artisan to examine the manufacture using means not directly applicable to examining the *Haemophilus influenzae* Rd genome or a subset thereof as it exists in nature or in purified form.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium which can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently know methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention.

A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and MicroSoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of dataprocessor structuring formats (e.g. text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing the nucleotide sequence of SEQ ID NO: 1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 in computer readable form, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al., J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al., Comp. Chem. 17*:203-207 (1993)) search algorithms on a Sybase system was used to identify open reading frames (ORFs) within the *Haemophilus influenzae* Rd genome which contain homology to ORFs or proteins from other organisms. Such ORFs are protein encoding fragments within the *Haemophilus influenzae* Rd genome and are useful in producing commercially important proteins such as enzymes used in fermentation reactions and in the production of commercially useful metabolites.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the *Haemophilus influenzae* Rd genome.

As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As stated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory which can store nucleotide sequence

information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention.

As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the *Haemophilus influenzae* Rd genome which match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are and can be used in the computer-based systems of the present invention. Examples of such software includes, but is not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA). A skilled artisan can readily recognize that any one of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that searches for commercially important fragments of the *Haemophilus influenzae* Rd genome, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymic active sites and signal sequences. Nucleic acid target motifs include,

but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the *Haemophilus influenzae* Rd genome possessing varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the *Haemophilus influenzae* Rd genome. In the present examples, implementing software which implement the BLAST and BLAZE algorithms (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) was used to identify open reading frames within the *Haemophilus influenzae* Rd genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

One application of this embodiment is provided in Figure 2. Figure 2 provides a block diagram of a computer system 102 that can be used to implement the present invention. The computer system 102 includes a processor 106 connected to a bus 104. Also connected to the bus 104 are a main memory 108 (preferably implemented as random access memory, RAM) and a variety of secondary storage devices 110, such as a hard drive 112 and a removable medium storage device 114. The removable medium storage device 114 may represent, for example, a floppy disk drive, a CD-ROM drive, a magnetic tape drive, etc. A removable storage medium 116 (such as a floppy disk, a compact disk, a magnetic tape, etc.) containing control logic and/or data recorded therein may be inserted into the removable medium storage device 114. The computer system 102 includes appropriate software

for reading the control logic and/or the data from the removable medium storage device 114 once inserted in the removable medium storage device 114.

A nucleotide sequence of the present invention may be stored in a well known manner in the main memory 108, any of the secondary storage devices 110, and/or a removable storage medium 116. Software for accessing and processing the genomic sequence (such as search tools, comparing tools, etc.) reside in main memory 108 during execution.

Biochemical Embodiments

Another embodiment of the present invention is directed to isolated fragments of the *Haemophilus influenzae* Rd genome. The fragments of the *Haemophilus influenzae* Rd genome of the present invention include, but are not limited to fragments which encode peptides, hereinafter open reading frames (ORFs), fragments which modulate the expression of an operably linked ORF, hereinafter expression modulating fragments (EMFs), fragments which mediate the uptake of a linked DNA fragment into a cell, hereinafter uptake modulating fragments (UMFs), and fragments which can be used to diagnose the presence of *Haemophilus influenzae* Rd in a sample, hereinafter diagnostic fragments (DFs).

As used herein, an "isolated nucleic acid molecule" or an "isolated fragment of the *Haemophilus influenzae* Rd genome" refers to a nucleic acid molecule possessing a specific nucleotide sequence which has been subjected to purification means to reduce, from the composition, the number of compounds which are normally associated with the composition. A variety of purification means can be used to generated the isolated fragments of the present invention. These include, but are not limited to methods which separate constituents of a solution based on charge, solubility, or size.

In one embodiment, *Haemophilus influenaze* Rd DNA can be mechanically sheared to produce fragments of 15-20 kb in length. These fragments can then be used to generate an *Haemophilus influenzae* Rd library

What Is Claimed Is:

1. Computer readable medium having recorded thereon the nucleotide sequence depicted in SEQ ID NO:1, a representative fragment thereof or a nucleotide sequence at least 99.9% identical to the nucleotide sequence depicted in SEQ ID NO:1.

2. Computer readable medium having recorded thereon any one of the fragments of SEQ ID NO:1 depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

3. The computer readable medium of claim 1, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

4. The computer readable medium of claim 3, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

5. A computer-based system for identifying fragments of the *Haemophilus* genome of commercial importance comprising the following elements;

a) a data storage means comprising the nucleotide sequence of SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1;

b) search means for comparing a target sequence to the nucleotide sequence of the data storage means of step (a) to identify homologous sequence(s), and

c) retrieval means for obtaining said homologous sequence(s) of step (b).

6. A method for identifying commercially important nucleic acid fragments of the *Haemophilus* genome comprising the step of comparing a database comprising the nucleotide sequence depicted in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 with a target sequence to obtain a nucleic acid molecule comprised of a complementary nucleotide sequence to said target sequence, wherein said target sequence is not randomly selected.

7. A method for identifying an expression modulating fragment of *Haemophilus* genome comprising the step of comparing a database comprising the nucleotide sequence depicted in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 with a target sequence to obtain a nucleic acid molecule comprised of a complementary nucleotide sequence to said target sequence, wherein said target sequence comprises sequences known to regulate gene expression.

8. An isolated protein-encoding nucleic acid fragment of the *Haemophilus influenzae* Rd genome, wherein said fragment consists of the nucleotide sequence of any one of the fragments of SEQ ID NO:1 depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

9. A vector comprising any one of the fragments of the *Haemophilus influenzae* Rd genome depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

10. An isolated fragment of the *Haemophilus influenzae* Rd genome, wherein said fragment modulates the expression of an operably linked

open reading frame, wherein said fragment consists of the nucleotide sequence from about 10 to 200 bases in length which is 5′ to any one of the open reading frames depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

11. A vector comprising any one of the fragments of the *Haemophilus influenzae* Rd genome of claim 8.

12. An organism which has been altered to contain any one of the fragments of the *Haemophilus* genome of claim 8.

13. An organism which has been altered to contain any one of the fragments of the *Haemophilus* genome of claim 10.

14. A method for regulating the expression of a nucleic acid molecule comprising the step of covalently attaching 5′ to said nucleic acid molecule a nucleic acid molecule consisting of the nucleotide sequence from about 10 to 100 bases 5′ to any one of the fragments of the *Haemophilus* genome depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

15. An isolated nucleic acid molecule encoding a homolog of any one of the fragment of the *Haemophilus* genome depicted in Table 1a, excluding the fragments of SEQ ID NO:1 depicted in Table 1b wherein said nucleic acid molecule is produced by the steps of:

a) screening a genomic DNA library using any one of the fragments of the *Haemophilus* genome depicted in Table 1a as a target sequence;

b) identifying members of said library which contain sequences which hybridize to said target sequence;

c) isolating the nucleic acid molecules from said members identified in step (b).

16. An isolated DNA molecule encoding a homolog of any one of the fragments of the *Haemophilus* genome depicted in Table 1a, excluding the fragments of SEQ ID NO:1 depicted in Table 1b wherein said nucleic acid molecule is produced by the steps of:

a) isolating mRNA, DNA, or cDNA produced from an organism;

b) amplifying nucleic acid molecules whose nucleotide sequence is homologous to amplification primers derived from said fragment of said *Haemophilus* genome to prime said amplification;

c) isolating said amplified sequences produced in step (b).

17. An isolated polypeptide encoded by any one of the fragments of the *Haemophilus influenzae* Rd genome depicted in Table 1a or by a degenerate variant of said fragment, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

18. An isolated polynucleotide molecule encoding any one of the polypeptides of claim 17.

19. An antibody which selectively binds to any one of the polypeptides of claim 17.

20. A method for producing a polypeptide in a host cell comprising the steps of:

a) incubating a host containing a heterologous nucleic acid molecule whose nucleotide sequence consists of any one of the fragments of the *Haemophilus influenzae* Rd genome depicted in Table 1a or a degenerate

variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b under conditions where said heterologous nucleic acid molecule is expressed to produce said protein, and

b) isolating said protein.